Novel machine-learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis

Grace Lai-Hung Wong, Vicki Wing-Ki Hui, Qingxiong Tan, Jingwen Xu, Hye Won Lee, Terry Cheuk-Fung Yip, Baoyao Yang, Yee-Kit Tse, Chong Yin, Fei Lyu, Jimmy Che-To Lai, Grace Chung-Yan Lui, Henry Lik-Yuen Chan, Pong-Chi Yuen, Vincent Wai-Sun Wong

Table of contents

Table S1. ICD-9-CM diagnosis and procedure codes for hepatic complications, liver cirrhosis, liver transplantation, HCC, and co-morbidities.

Disease	ICD-9-CM Code	Description
Hepatitis		
Chronic hepatitis B	070.22	Chronic viral hepatitis B with hepatic coma without hepatitis delta
Chronic hepatitis B	070.23	Chronic viral hepatitis B with hepatic coma with hepatitis delta
Chronic hepatitis B	070.32	Chronic viral hepatitis B without mention of hepatic coma without mention of hepatitis delta
Chronic hepatitis B	070.33	Chronic viral hepatitis B without mention of hepatic coma with hepatitis delta
Chronic hepatitis B	V02.61	Hepatitis B carrier
Acute hepatitis B	070.20	Viral hepatitis B with hepatic coma, acute or unspecified, without mention of hepatitis delta
Acute hepatitis B	070.21	Viral hepatitis B with hepatic coma, acute or unspecified, with hepatitis delta
Acute hepatitis B	070.30	Viral hepatitis B without mention of hepatic coma, acute or unspecified, without mention of hepatitis delta
Acute hepatitis B	070.31	Viral hepatitis B without mention of hepatic coma, acute or unspecified, with hepatitis delta
Hepatitis C	070.41	Acute hepatitis C with hepatic coma
Hepatitis C	070.44	Chronic hepatitis C with hepatic coma
Hepatitis C	070.51	Acute hepatitis C without mention of hepatic coma
Hepatitis C	070.54	Chronic hepatitis C without mention of hepatic coma
Hepatitis C	V02.62	Hepatitis C carrier
Hepatitis D	070.23	Chronic viral hepatitis B with hepatic coma with hepatitis delta
Hepatitis D	070.33	Chronic viral hepatitis B without mention of hepatic coma with hepatitis delta
Hepatitis D	070.42	Hepatitis delta without mention of active hepatitis B disease with hepatic coma
Hepatitis D	070.52	Hepatitis delta without mention of active hepatitis B disease or hepatic coma
Human immunodeficiency virus (HIV)		
HIV	042	Human immunodeficiency virus [HIV] disease.
HIV	079.53	Human immunodeficiency virus, type 2 [HIV-2]
HIV	V02.9:1	HIV carrier
HIV	V08	Asymptomatic human immunodeficiency virus [HIV] infection status
Hepatic complications		
Ascites	789.5	Ascites
SBP	567.2:9	Spontaneous bacterial peritonitis
SBP	567.8:0	Peritonitis
EVB[#]	456.0	Esophageal varices with bleeding
EVB	456.20	Esophageal varices classified elsewhere with bleeding
GVB[#]	456.8:1	Fundal varices, bleeding
GVB	456.8:2	Bleeding gastric varices
HE	348.3	Encephalopathy, unspecified
HE	349.82	Toxic encephalopathy
HE	572.2	Hepatic coma

HRS	572.4	Hepatorenal syndrome
Portal hypertension	572.3	Portal hypertension
Varices	456.1	Esophageal varices without bleeding
Varices	456.21	Esophageal varices in diseases classified elsewhere without bleeding
Varices	456.8:4	Fundal varices
Varices	456.8:5	Gastric varices
Liver cirrhosis		
Liver cirrhosis	571.2	Alcoholic cirrhosis of liver
Liver cirrhosis	571.5	Cirrhosis of liver without mention of alcohol
Liver transplantation		
Liver transplantation	V42.7	Liver replaced by transplant
Liver transplantation	50.51	Auxiliary liver transplant
Liver transplantation	50.59	Other transplant of liver
Hepatocellular carcinoma (HCC)		
HCC	155.0	Malignant neoplasm of liver, primary
HCC	155.2	Malignant neoplasm of liver, not specified
HCC*	197.7	Secondary malignant neoplasm of liver
Co-morbidities		
Hypertension	401	Essential hypertension
Hypertension	401.0	Malignant essential hypertension
Hypertension	401.1	Benign essential hypertension
Hypertension	401.9	Unspecified essential hypertension
Chronic kidney disease	585.1-585.9, 586	Chronic kidney disease
Hyperlipidemia	272.0-272.9	Hyperlipidemia, Hypercholesterolemia
Osteopenia	733.9	Osteopenia
Osteoporosis	733.00-733.01	Osteoporosis
Neoplasms	140 to 239	Neoplasms, malignancies, cancers
Mental disorders	280 to 289	Mental disorders
Diabetes mellitus	250.0-250.9	Diabetes mellitus
Cardiovascular disease	390 to 459	Cardiovascular disease

* ICD-9-CM diagnosis code 197.7 was treated as primary liver cancer if there was no other primary cancer coded; the definition of HCC also considered the ICD-9-CM procedure codes of 50.22, 50.99, 50.29, 50.94, 38.80, and 88.47

[#] Esophageal or gastric variceal bleeding was also defined by the ICD-9-CM procedure codes of 42.33:3, 42.33:6, 42.33:13, and 43.41:1

Abbreviations: EVB = esophageal variceal bleeding; GVB = gastric variceal bleeding, HCC = hepatocellular carcinoma, HE = hepatic encephalopathy, HIV = human immunodeficiency virus, HRS = hepatorenal syndrome, ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification, SBP = spontaneous bacterial peritonitis.

Table S2. Drug codes of nucleos(t)ide analogues and other concomitant medications used in Hospital Authority internally.

Drug code	Name	Dosage
Antiviral treatment		
ADEF01	Adefovir Dipivoxil	10 MG
ENTE01/02	Entecavir	0.5 MG/1.0 MG
INTE04/05/18/19	Interferon alpha-2a	3-9MIU/0.5-1ML
INTE06-09/16-17	Interferon alpha-2b	3/5/10/15/25 MIU/1ML
LAMI07/08/09/10	Lamivudine	150 MG/10 MG/ML/100MG
PEGI01-03/05/09-12/18-21	Peginterferon alpha-2b	50-120 MCG/0.5-1ML
PEGI04/06-08/13/15-17	Peginterferon alpha-2a	135-180 MCG/0.5-1ML
PEGI14	Peginterferon lambda-1a	180 MCG/0.45ML
TELB01	Telbivudine	600 MG
TENO03/04/07	Tenofovir Disoproxil Fumarate	300 MG
TENO06/08	Tenofovir Alafenamide	25 MG
Anti-diabetic agents		
ACAR01/02	Acarbose	100 MG / 50 MG
ACTO09	Actosmet	15 MG / 850 MG
ALOG01/02	Alogliptin	25 MG / 12.5 MG
AVAN03/04	Avadma,et	2 or 4 MG / 1000 MG
S01057	Canagliflozin	100 MG
CHLO51	Chlorpropamide	250 MG
DAPA01	Dapagliflozin	10 MG
EMPA01/02	Empagliflozin	10 MG / 25 MG
EXEN03/04/05	Exenatide	5 MCG / 10 MCG
GALV01/02/03	Galvus Met	50 MG / 500 or 1000 or 850 MG
GLIB01/02	Glibenclamide	2.5 MG / 5 MG
GLIC01/02/03	Gliclazide	80 MG / 30 MG / 60 MG
GLIP01	Glipizide	5 MG
JANU01/02	Janumet	50 MG / 500 or 1000 MG
LINA01	Linagliptin	5 MG
LIRA01/02	Liraglutide	6 MG/ML
LIXI01/02	Lixisenatide	10 MCG / 20 MCG
METF01/02/03/05	Metformin	250 MG / 500 MG / 1000 MG
OSEN01/02	Oseni	25 MG / 15 or 30 MG
PIOG01/02	Pioglitazone	30 MG / 15 MG
ROSI01/02/03	Rosiglitazone	2 MG / 4 MG / 8 MG
SITA02/03/06	Sitagliptin	50 MG / 100 MG / 25 MG
TOLB01	Tolbutamide	500 MG
TRAJ01/02	Trajenta Duo	2.5 MG / 500 or 1000 MG
VILD01	Vildagliptin	50 MG
INSU01-53	Insulin	100U / ML
Lipid-lowering agents		
ATOR01/02/03/04	Atorvastatin	10 MG / 20 MG / 40 MG / 80 MG
FLUV02/03/05	Fluvastatin	20 MG / 40 MG / 80 MG
LOVA01	Lovastatin	20 MG
PRAV01/02	Pravastatin	10 MG / 20 MG

ROSU01/02	Rosuvastatin	10 MG / 20 MG
SIMV01/02/04/05	Simvastatom	10 MG / 20 MG / 40 MG / 80 MG
BEZA01/02	Bezafibrate	200 MG / 400 MG
EXE01	Ezetimibe	10 MG
FENO01/05/06/07	Fenofibrate	100 MG / 200 MG /160 MG / 145 MG
GEMF01/02/03	Gemfibrozil	300 MG / 600 MG / 900 MG
Anti-hypertensive agents		
AMLO01/02/03	Amlodipine	10 MG / 5 MG
ATEN01/02	Atenolol	50 MG / 100 MG
CADN01/02	Candesartan	8 MG / 16 MG
CAPT01/02/03/04/06	Captopril	25 MG / 50 MG / 12.5 MG / 6.25 MG
CARV01/02/03/04	Caredilol	25 MG / 12.5 MG / 6.25 MG / 3.125 MG
CO-D01/02	Co-Diovan	80 or 160 MG / 12.5 MG
DILT01-08	Diltiazem	25 MG – 200 MG
DOXA03/05/06	Doxazosin	2 MG / 4 MG / 8 MG
ENAL01/02/03/04	Enalapril	20 MG / 10 MG / 5 MG / 1 MG
FEL01/02/03	Felodipine	5 MG / 10 MG / 2.5 MG
HYDR01/02/03/65	Hydralazine	10 MG / 25 MG / 50 MG / 20 MG
HYDR05/30/38	Hydrochlorothiazide	20 MG / 25 MG / 2 MG/ML
INDA01/02	Indapamide	2.5 MG / 1.5 MG
IRBE01/02/03/04	Irbesartan	150 MG / 300 MG
LABE01/02/03/04/05	Labetalol	100 MG / 200 MG / 50 MG
LERC01	Lercanidipine	10 MG
LISI01/02/03	Lisinopril	5 MG / 10 MG / 20 MG
LOSA01/02/03/04	Losartan	50 MG / 100 MG
METH22/23/78	Methyldopa	125 MG / 250 MG / 50 MG/ML
METO09-17	Metoprolol	100 MG / 50 MG / 25 MG
NIFE01-05	Nifedipine	5 MG / 20 MG / 30 MG / 60 MG
NIMO01/02	Nimodipine	30 MG / 10 MG/50ML
PERI17/28/29	Perinopril	2 MG / 4 MG / 5 MG
PRAZ03/04/05	Prazosin	1 MG / 2 MG / 5 MG
PROP04-30	Propranolol	1 MG – 160 MG
RAMI01/02	Ramipril	2.5 MG / 5 MG
TELM01/02	Telmisartan	40 MG / 80 MG
VALS02/03	Valsartan	80 MG / 160 MG
VERA01/02/03/04	Verapamil	40 MG / 80 MG / 240 MG / 5MG/2ML

Table S3. List of viral serological markers retrieved.

HBV	HCV	HDV
Anti-HBc	Anti-HCV	Anti-HDV
Anti-HBc IgM	HCV RNA (viral load), RT-	
Anti-HBe	HCV RNA, RT-PCR	
Anti-HBs		
Anti-HBs, Quantitative		
HBeAg		
HBsAg		
HBV DNA		
HBV DNA (viral load), RT-PCR		

Anti-HBc = antibody to hepatitis B core antigen; Anti-HBe = antibody to hepatitis B e antigen; Anti-HBs = antibody to hepatitis B surface antigen; HBeAg = hepatitis B e antigen; HBsAg = hepatitis B surface antigen; HBV = hepatitis B virus; HCV = hepatitis C virus; HDV = hepatitis D virus; IgM = immunoglobulin M; RT-PCR = Reverse transcription polymerase chain reaction.

Table S4. Serum test formulae for liver fibrosis.

PARAMETERS OR INDEX	FORMULA
APRI	$\dfrac{\dfrac{AST\ level\ of\ patient}{AST\ upper\ limit\ of\ normal}}{Platelet\ count\ (10^9/L)} \times\ 100\%$
FORNS INDEX	$7.811 - 3.131 \times \ln\,(platelet\ count\ (10^9/L)) + 0.781 \times \ln\big(GGT(IU/L)\big) + 3.467 \times \ln\ (age) - 0.014 \times cholesterol\ (mg/dL)$
FIB-4	$\dfrac{Age\ (years) \times AST(U/L)}{Platelet\ count\ (10^9/L) \times \sqrt{ALT\ (U/L)}}$

Table S5. The detailed parameters for the machine learning models

Machine learning model	Detailed model parameters	
Logistic regression	**Maximum number of iterations**	100
	Optimization Algorithm	“Newton-CG” algorithm
$^{+-}$Ridge regression	**Maximum number of iterations**	1000
AdaBoost	**Base estimator** （The base estimator from which the boosted ensemble is built）	Decision Tree Classifier with max depth of 1
	The number of estimators (The maximum number of estimators at which boosting is terminate）	50
	Learning rate	1.0
$^{+}$Decision tree	**Max depth** (The maximum depth of the tree)	10
	Criterion (The function to measure the quality of a split)	Gini Impurity
	Max features (The number of features to consider when looking for the best split)	Select parameter number (20,36,or all)
	Splitter (The strategy used to choose the split at each node)	Use the “best” strategy to choose the best split
	minimum samples of each leaf (The minimum number of samples required to be at a leaf node)	1
$^{+-}$Random Forest	**The number of trees in the forest**	20
	Max depth (The maximum depth of the tree)	10
	Max features (The number of features to consider when looking for the best split)	Select parameter number (20,36,or all)
	Samples of each leaf (The minimum number of samples required to be at a leaf node)	1
	The minimum samples for split (The minimum number of samples required to split an internal node)	2

Reference:

[1] Hilbe JM. Logistic regression models. Boca Raton, Florida, USA: CRC press; 2009.

[2] Montgomery DC, Peckl EA, Vining GG. Introduction to linear regression analysis. Hoboken, New Jersey, USA: John Wiley & Sons; 2015.

[3] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. In: Dresher M, Tucker AW, Wolfe P, eds. European conference on computational learning theory. Berlin/Heidelberg, Germany: Springer Berlin Heidelberg; 1995:23-37.

[4] Breiman L, Friedman J, Stone CJ, Olsehn RA. Classification and regression trees. CRC Press; 1984.

[5] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003 Nov 24;43(6):1947-58.

[6] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., and Duchesnay E. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011, 12, 2825-2830.

Table S6. Antiviral treatment received by patients cumulated over the four periods.

Period	2000 – 2004	2005 – 2009	2010 – 2013	2014 – 2018
Antiviral treatment for chronic hepatitis B (n, %)				
Lamivudine	2,059(10.80%)	4,934(10.10%)	7,985(9.30%)	9,464(7.46%)
Adefovir Dipivoxil	38(0.20%)	1,180(2.41%)	2,611(3.04%)	2,954(2.33%)
Entecavir	0(0.00%)	1,214(2.48%)	7,709(8.98%)	42,494(33.49%)
Telbivudine	0(0.00%)	229(0.47%)	1,443(1.68%)	2,741(2.16%)
Tenofovir*	3(0.02%)	102(0.21%)	1,007(1.17%)	6,036(4.76%)
Any nucleos(t)ide analogues	2,077(10.90%)	6,642(13.59%)	16,512(19.23%)	51,191(40.34%)
Conventional interferon	99(0.52%)	53(0.11%)	63(0.07%)	72(0.06%)
Peginterferon	11(0.06%)	375(0.77%)	690(0.80%)	909(0.72%)
Any antiviral treatment	2,248(11.79%)	5,889(12.05%)	15,249(17.76%)	51,572(40.64%)
Cumulative no. of CHB patients	**19,060**	**48,869**	**85,880**	**126,890**
Antiviral treatment for chronic hepatitis C (n, %)				
Conventional or pegylated interferon + ribavirin	1,593(29.71%)	2,754(30.41%)	3,926(31.84%)	5,219(31.05%)
Sofosbuvir/velpatasvir	0(0.00%)	0(0.00%)	0(0.00%)	273(1.62%)
Sofosbuvir/ledipasvir	0(0.00%)	0(0.00%)	0(0.00%)	12(0.07%)
Dasabuvir/Ombitasvir/paritaprevir combination therapy	0(0.00%)	0(0.00%)	0(0.00%)	119(0.71%)
Elbasvir/Grazoprevir	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
Asunaprevir+/-Daclatasvir	0(0.00%)	0(0.00%)	0(0.00%)	107(0.64%)
Glecaprevir/Pibrentasvir	0(0.00%)	0(0.00%)	0(0.00%)	6(0.04%)
Boceprevir	0(0.00%)	0(0.00%)	1(0.01%)	3(0.01%)
Any antiviral treatment	1,593(29.71%)	2,754(30.41%)	3,927(31.84%)	5,660(33.67%)
Cumulative no. of CHC patients	**5,362**	**9,056**	**12,335**	**16,811**

Descriptive statistics were calculated after subtraction of missing data from denominator.

CHB = chronic hepatitis B ; CHC = chronic hepatitis C; NA = nucleos(t)ide analogues.

Table S7. Untreated Patients with advanced fibrosis during follow-up based on serum formulae – current situation and projected disease burden (only in patients with results of at least one of the three serum fibrosis scores)

Period	2000 – 2004	2005 – 2009	2010 – 2013	2014 – 2018
	Patients with chronic hepatitis B (N = 44,193)			
No. of subjects	**N = 3,032**	**N = 11,143**	**N = 13,550**	**N = 16,468**
APRI	1.90(6.28)	1.02(3.45)	0.93(2.64)	1.10(6.12)
	0.63[0.,1.56]	0.43[0.,0.90]	0.39[0.,0.75]	0.38[0.,0.75]
Missing (%)	214(7.06%)	835(7.49%)	1577(11.64%)	1932(11.73%)
Forns index	6.05(2.70)	5.83(2.34)	5.95(2.21)	6.32(2.34)
Missing (%)	2305(76.02%)	7034(63.12%)	8230(60.74%)	9163(55.64%)
FIB-4	0.75(1.98)	0.55(1.93)	0.62(2.05)	0.84(4.53)
	0.30[0.,0.72]	0.26[0.,0.51]	0.29[0.,0.53]	0.32[0.,0.61]
Missing (%)	215(7.09%)	835(7.49%)	1579(11.65%)	1933(11.74%)
Advanced liver fibrosis	662(21.83%)	1444(12.96%)	1499(11.06%)	2244(13.63%)
APRI ≥ 2	552(19.59%)	1017(9.87%)	948(7.92%)	1307(8.99%)
FIB-4 ≥ 3.25	98(3.48%)	176(1.71%)	246(2.05%)	447(3.08%)
Forns index ≥ 8.4	163(22.42%)	609(14.82%)	725(13.63%)	1211(16.58%)
ALT > 2xULN	759(25.04%)	1954(17.54%)	1732(12.78%)	2075(12.60%)
Missing (%)	1(0.03%)	0(0.00%)	2(0.01%)	0(0.00%)
	Patients with chronic hepatitis C (N = 5,249)			
No. of subjects	**N = 1,105**	**N = 1,357**	**N = 1,272**	**N = 1,515**
APRI	1.45(5.27)	1.26(2.71)	1.19(2.03)	1.17(3.97)
	0.56[0.,1.32]	0.56[0.,1.15]	0.54[0.,1.19]	0.49[0.,0.95]
Missing (%)	34(3.08%)	99(7.30%)	121(9.51%)	155(10.23%)
Forns index	7.30(2.45)	7.08(2.42)	7.01(2.47)	6.56(2.26)
Missing (%)	962(87.06%)	995(73.32%)	880(69.18%)	983(64.88%)
FIB-4	0.64(1.29)	0.72(1.70)	0.75(1.71)	0.83(4.31)
	0.34[0.,0.67]	0.34[0.,0.71]	0.35[0.,0.70]	0.33[0.,0.63]
Missing (%)	34(3.08%)	99(7.30%)	121(9.51%)	155(10.23%)
Advanced liver fibrosis	206(18.64%)	239(17.61%)	252(19.81%)	206(13.60%)
APRI ≥ 2	174(16.25%)	172(13.67%)	168(14.60%)	131(9.63%)
FIB-4 ≥ 3.25	21(1.96%)	40(3.18%)	44(3.82%)	39(2.87%)
Forns index ≥ 8.4	50(34.97%)	101(27.90%)	114(29.08%)	99(18.61%)
ALT > 2xULN	241(21.81%)	266(19.60%)	254(19.97%)	266(17.56%)
Missing (%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)

APRI = AST to Platelet Ratio Index.

Table S8. Diagnostic accuracy of the HCC-RS using selected parameters at different cut-offs in the training (n=86,804) and validation (n=37,202) cohorts.

Cohort	Cut-off	n (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
		(≥ cut-off)	(95% CI)	(95% CI)	(95% CI)	(95% CI)
Training cohort	0.1	21,652 (24.94)	0.7276 [0.7165-0.7382]	0.7913 [0.7886-0.7942]	0.2292 [0.2238-0.2350]	0.9715 [0.9703-0.9728]
	0.2	6,315 (7.28)	0.3466 [0.3347-0.3578]	0.9506 [0.9491-0.9521]	0.3743 [0.3622-0.3864]	0.9446 [0.9429-0.9462]
	0.3	2,158 (2.49)	0.1182 [0.1105-0.1260]	0.9831 [0.9822-0.9839]	0.3735 [0.3523-0.3954]	0.9289 [0.9272-0.9307]
	0.4	329 (0.38)	0.0172 [0.0143--0.0203]	0.9973 [0.9970-0.9977]	0.3556 [0.3055-0.4091]	0.9225 [0.9207-0.9243]
	0.5	2 (0.00)	0.0000 [0.0000-0.0000]	1.0000 [0.9999-1.0000]	0.0000 [0.0000-0.0000]	0.9214 [0.9203-0.9239]
	0.6	0 (0.00)	0.0000 [0.0000-0.0000]	1.0000 [1.0000-1.0000]	nan	0.9214 [0.9196-0.9232]
	0.7	0 (0.00)	0.0000 [0.0000-0.0000]	1.0000 [1.0000-1.0000]	nan	0.9214 [0.9196-0.9232]
Validation cohort	0.1	9,216 (24.77)	0.7263 [0.7092-0.7420]	0.7924 [0.7882-0.7965]	0.2266 [0.2176-0.2355]	0.9719 [0.9698-0.9738]
	0.2	2,622 (7.05)	0.3492 [0.3309-0.3667]	0.9529 [0.9507-0.9552]	0.3829 [0.3644-0.4022]	0.9459 [0.9434-0.9483]
	0.3	932 (2.51)	0.1343 [0.1221-0.1468]	0.9841 [0.9828-0.9854]	0.4142 [0.3839-0.4457]	0.9314 [0.9288-0.9342]
	0.4	122 (0.33)	0.0167 [0.0120-0.0216]	0.9978 [0.9974-0.9983]	0.3934 [0.3143-0.4806]	0.9238 [0.9211-0.9267]
	0.5	1 (0.00)	0.0003 [0.0003-0.0004]	1.0000 [1.0000-1.0000]	1.0000 [1.0000-1.0000]	0.9227 [0.9205-0.9251]
	0.6	0 (0.00)	0.0000 [0.0000-0.0000]	1.0000 [1.0000-1.0000]	nan	0.9227 [0.9200-0.9254]
	0.7	0 (0.00)	0.0000 [0.0000-0.0000]	1.0000 [1.0000-1.0000]	nan	0.9227 [0.9200-0.9254]

CI=confidence interval, NPV=negative predictive value, PPV=positive predictive value

Table S9. Area under the receiver operating characteristic curve (AUROC) and the 95% confidence interval of the machine learning models in the training cohort and external validation cohort of 4,462 Korean patients.

Machine learning model	*Training cohort: N = 86,804, HCC = 6,821	^Validation cohort: N =4,462, HCC = 1,072
	Selected parameters (num=20)	*Selected parameters (num=20)*
Logistic regression	0.814±0.006	0.813±0.009
Ridge regression	0.817±0.006	0.819±0.006
AdaBoost	0.822±0.006	0.821±0.009
Decision tree	0.877±0.005	0.799±0.030
-Random Forest	0.987±0.003	0.801±0.030

*AUROC of the five machine learning algorithms were overall difference in training cohort, *P*<0.05.

^AUROC of the five machine learning algorithms were overall difference in validation cohort, *P*<0.05.

- AUROC higher than decision tree in validation cohort, *P*<0.05

Table S10. Accuracy of the machine learning models using selected parameters in diagnosing HCC in the external validation cohort of 4,462 Korean patients.

Machine learning algorithm	Dual Cut-offs	n (%) (< lower cut-off / ≥ upper cut-off)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)
Logistic regression	0.21	2,215 (49.7)	0.90 (0.89-0.91)	0.542 (0.539-0.545)	0.137 (0.134-0.140)	0.973 (0.969-0.977)
	0.32	571 (12.8)	0.42 (0.39-0.45)	0.900 (0.896-0.902)	0.276 (0.273-0.279)	0.948 (0.944-0.952)
Ridge regression	0.08	2,293 (51.4)	0.90 (0.89-0.91)	0.587 (0.583-0.591)	0.154 (0.150-0.158)	0.983 (0.980-0.986)
	0.16	597 (13.4)	0.41 (0.37-0.45)	0.900 (0.897-0.903)	0.268 (0.264-0.272)	0.949 (0.945-0.953)
AdaBoost	0.37	2,284 (51.2)	0.91 (0.90-0.92)	0.529 (0.525-0.534)	0.139 (0.134-0.142)	0.976 (0.973-0.979)
	0.41	588 (13.2)	0.41 (0.39-0.43)	0.908 (0.904-0.912)	0.297 (0.295-0.299)	0.951 (0.947-0.955)
Decision tree	0.04	2,065 (46.3)	0.90 (0.89-0.91)	0.498 (0.496-0.501)	0.129 (0.125-0.133)	0.979 (0.976-0.982)
	0.19	566 (12.7)	0.39 (0.37-0.41)	0.900 (0.897-0.904)	0.298 (0.294-0.302)	0.948 (0.941-0.955)
Random forest	0.02	2,208 (49.5)	0.90 (0.89-0.91)	0.497 (0.491-0.503)	0.128 (0.124-0.132)	0.976 (0.971-0.981)
	0.22	499 (11.2)	0.38 (0.36-0.40)	0.910 (0.907-0.913)	0.283 (0.280-0.286)	0.939 (0.936-0.942)

CI=confidence interval, NPV=negative predictive value, PPV=positive predictive value.

Table S11. Area under the receiver operating characteristic curve (AUROC) and the 95% confidence interval of the machine learning models in training and validation cohorts to HCC, with dataset in 2000-2010 and 2011-2018 respectively.

MACHINE LEARNING MODEL	*TRAINING COHORT: N = 42,254, HCC = 3,231			^VALIDATION COHORT: N = 18,109, HCC = 1,404		
	20 selected parameters	*36 selected parameters*	*All parameters*	*20 selected parameters*	*36 selected parameters*	*All parameters*
	Data set in 2000-2010					
LOGISTIC REGRESSION	0.826±0.006	0.830±0.006	0.835±0.006	0.827±0.009	0.829±0.009	0.842±0.009
RIDGE REGRESSION	0.827±0.005	0.835±0.005	0.847±0.005	0.828±0.009	0.836±0.009	0.854±0.009
ADABOOST	0.829±0.006	0.829±0.006	0.836±0.006	0.826±0.009	0.826±0.009	0.841±0.009
DECISION TREE	0.892±0.005	0.894±0.005	0.896±0.005	0.784±0.010	0.792±0.010	0.801±0.010
-RANDOM FOREST	0.984±0.003	0.988±0.003	0.988±0.003	0.797±0.010	0.806±0.010	0.824±0.010
	Data set in 2011-2019					
LOGISTIC REGRESSION	0.811±0.006	0.819±0.006	0.818±0.006	0.802±0.009	0.811±0.009	0.815±0.009
-RIDGE REGRESSION	0.816±0.005	0.830±0.005	0.842±0.006	0.808±0.009	0.823±0.008	0.836±0.008
ADABOOST	0.824±0.006	0.831±0.006	0.832±0.006	0.812±0.009	0.817±0.009	0.821±0.009
DECISION TREE	0.884±0.005	0.888±0.005	0.893±0.005	0.776±0.010	0.778±0.010	0.797±0.010
-RANDOM FOREST	0.992±0.003	0.995±0.003	0.996±0.003	0.776±0.010	0.806±0.010	0.802±0.010

*AUROC of the five machine learning algorithms were overall difference in training cohort, *P*<0.05.

^AUROC of the five machine learning algorithms were overall difference in validation cohort, *P*<0.05.

- AUROC higher than decision tree in validation cohort, *P*<0.05.

Table S12. Accuracy of the machine learning models using selected parameters in diagnosing HCC in the training and validation cohorts. In the training cohort, dual cut-offs were selected to achieve >90% sensitivity and specificity, with dataset in 2000-2010.

Machine learning algorithm	Dual Cut-offs	n (%) (< lower cut-off / ≥ upper cut-off)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)
Training cohort (n=42,254)						
Logistic regression	0.17	23,200 (54.9)	0.90 (0.89-0.91)	0.586 (0.584-0.588)	0.152 (0.151-0.153)	0.986 (0.984-0.988)
	0.29	5,506 (13.0)	0.49 (0.48-0.50)	0.90 0(0.898-0.902)	0.291 (0.289-0.293)	0.955 (0.953-0.957)
Ridge regression	0.06	24,253 (57.3)	0.90 (0.89-0.91)	0.613 (0.610-0.616)	0.161 (0.158-0.164)	0.986 (0.985-0.987)
	0.15	5,641 (13.4)	0.53(0.52-0.54)	0.900 (0.898-0.902)	0.308 (0.304-0.312)	0.959 (0.955-0.958)
AdaBoost	0.42	22,706 (53.7)	0.90 (0.89-0.91)	0.573 (0.570-0.576)	0.149 (0.147-0.151)	0.986 (0.985-0.987)
	0.46	5,532 (13.1)	0.52 (0.51-0.53)	0.901 (0.899-0.903)	0.305 (0.303-0.307)	0.958 (0.957-0.959)
Decision tree	0.04	26,618 (63.0)	0.91 (0.90-0.92)	0.675 (0.672-0.678)	0.189 (0.187-0.191)	0.989 (0.988-0.990)
	0.17	6,019 (14.2)	0.68 (0.66-0.70)	0.902 (0.901-0.903)	0.366(0.364-0.368)	0.971 (0.969-0.973)
Random forest	0.35	39,103 (92.5)	0.90 (0.90-0.91)	0.993 (0.992-0.994)	0.923 (0.921-0.925)	0.991 (0.989-0.993)
	0.10	6,303 (14.9)	0.95(0.94-0.96)	0.917 (0.915-0.918)	0.491 (0.489-0.493)	0.996 (0.994-0.998)
Validation cohort (n=18,109)						
Logistic regression	0.18	10,204 (56.3)	0.90 (0.89-0.91)	0.602 (0.599-0.605)	0.159 (0.156-0.162)	0.986 (0.984-0.988)
	0.30	2,375 (13.1)	0.50(0.48-0.52)	0.900 (0.896-0.902)	0.298 (0.295-0.301)	0.955 (0.953-0.957)
Ridge regression	0.06	10,877 (60.1)	0.90 (0.89-0.91)	0.642 (0.638-0.646)	0.174 (0.170-0.178)	0.987 (0.985-0.989)
	0.15	2,439 (13.5)	0.54 (0.52-0.56)	0.900 (0.898-0.902)	0.315 (0.311-0.319)	0.959 (0.956-0.962)
AdaBoost	0.42	9,771 (54.1)	0.90 (0.89-0.91)	0.576 (0.574-0.578)	0.152(0.149-0.155)	0.986 (0.984-0.988)
	0.46	2,372 (13.1)	0.53(0.51-0.55)	0.903 (0.901-0.905)	0.317 (0.314-0.320)	0.958 (0.956-0.960)
Decision tree	0.02	8,200 (45.1)	0.90 (0.89-0.91)	0.480 (0.478-0.482)	0.123 (0.121-0.125)	0.978 (0.975-0.981)
	0.19	2,047 (11.3)	0.46 (0.52-0.56)	0.906 (0.904-0.908)	0.320 (0.318-0.322)	0.953 (0.951-0.955)
Random forest	0.01	9,133 (50.4)	0.90 (0.89-0.91)	0.538 (0.536-0.540)	0.140 (0.138-0.142)	0.984 (0.982-0.986)
	0.25	2,084 (11.5)	0.46 (0.44-0.48)	0.914 (0.911-0.917)	0.316 (0.310-0.322)	0.953 (0.950-0.956)

CI=confidence interval, NPV=negative predictive value, PPV=positive predictive value.

Table S13. Accuracy of the machine learning models using selected parameters in diagnosing HCC in the training and validation cohorts. In the training cohort, dual cut-offs were selected to achieve >90% sensitivity and specificity, with dataset in 2011-2019.

Machine learning algorithm	Dual Cut-offs	n (%) (< lower cut-off / ≥ upper cut-off)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)
			Training cohort (n=44,551)			
Logistic regression	0.18	21,658 (48.6)	0.90 (0.89-0.91)	0.519 (0.516-0.522)	0.139 (0.136-0.142)	0.983 (0.980-0.986)
	0.29	5,894(13.2)	0.50 (0.49-0.51)	0.90 (0.898-0.902)	0.304 (0.300-0.308)	0.954 (0.951-0.957)
Ridge regression	0.06	24,212 (54.3)	0.90 (0.89-0.91)	0.581 (0.578-0.584)	0.156 (0.152-0.160)	0.983 (0.980-0.986)
	0.14	6,022 (13.5)	0.54(0.51-0.57)	0.900 (0.898-0.902)	0.319 (0.316-0.322)	0.958 (0.956-0.960)
AdaBoost	0.42	22,283 (50.0)	0.91 (0.90-0.92)	0.535 (0.532-0.538)	0.145 (0.142-0.148)	0.986 (0.984-0.988)
	0.45	5,067 (11.4)	0.47 (0.45-0.49)	0.917 (0.911-0.923)	0.331 (0.328-0.334)	0.952 (0.949-0.955)
Decision tree	0.03	27,733 (62.2)	0.90 (0.89-0.91)	0.667 (0.664-0.670)	0.189 (0.185-0.193)	0.987 (0.985-0.989)
	0.13	5,465 (12.3)	0.63 (0.62-0.64)	0.915 (0.913-0.917)	0.412(0.409-0.415)	0.967 (0.964-0.970)
Random forest	0.55	41,281 (92.7)	0.91(0.89-0.93)	0.997 (0.996-0.998)	0.991 (0.989-0.993)	0.992 (0.991-0.993)
	0.10	6,352 (14.3)	0.98 (0.97-0.98)	0.930 (0.928-0.932)	0.548 (0.541-0.555)	0.997 (0.997-0.998)
			Validation cohort (n=19,094)			
Logistic regression	0.17	8,961 (46.9)	0.90 (0.89-0.91)	0.501 (0.498-0.504)	0.135 (0.132-0.138)	0.983 (0.980-0.988)
	0.29	2,519 (13.2)	0.50 (0.58-0.52)	0.900 (0.897-0.903)	0.302 (0.298-0.306)	0.954 (0.951-0.957)
Ridge regression	0.06	10,190 (53.3)	0.90 (0.89-0.91)	0.571 (0.567-0.575)	0.153 (0.150-0.156)	0.985 (0.983-0.987)
	0.14	2,560 (13.4)	0.53 (0.50-0.56)	0.900 (0.897-0.903)	0.316 (0.311-0.321)	0.956 (0.951-0.960)
AdaBoost	0.41	9,626 (50.4)	0.90(0.88-0.92)	0.539 (0.536-0.542)	0.145 (0.142-0.148)	0.984 (0.981-0.987)
	0.45	2,137 (11.2)	0.45 (0.43-0.47)	0.915 (0.913-0.917)	0.319 (0.317-0.322)	0.951 (0.949-0.953)
Decision tree	0.01	7,563 (39.6)	0.90 (0.89-0.91)	0.45 3(0.451-0.455)	0.125 (0.122-0.128)	0.975 (0.971-0.979)
	0.13	2,327 (12.2)	0.48 (0.45-0.51)	0.909 (0.904-0.914)	0.317 (0.314-0.320)	0.953 (0.950-0.956)
Random forest	0.01	9,494 (49.7)	0.90 (0.89-0.91)	0.520 (0.517-0.523)	0.137 (0.132-0.142)	0.978 (0.974-0.982)
	0.25	2,049 (10.7)	0.46 (0.42-0.50)	0.923 (0.919-0.927)	0.342 (0.338-0.346)	0.951 (0.948-0.954)

CI=confidence interval, NPV=negative predictive value, PPV=positive predictive value.